Konstruktie en scoringssysteem van de I.S.I.-interessetest, vorm I en II

Welten, Vincentius Johannes

Summary[1]

In the first part of this study an account is given of the construction and scoring system of the ISI Interest test, Form I.

It seems at present that interests can best be investigated by means of an interest-inventory, in which the subject states his preference for or aversion to a large number of things or activities. These subjective self-descriptions can be summarised in two different ways: in the one, the various activities of the test are combined in more homogeneous groups, so that scores can be obtained for a number of divergent fields of interest; in the other, items are incorporated in so-called differentiating scales.[2] These differentiating scales do not represent particular fields of interest, but facilitate maximal differentiation between certain groups of people.

Homogeneous scales alone have been developed for Form I of the interest test. The problems attached to this may be set in the following separate questions:

a. Have children of 11–13 years of age a consistent preference for particular fields of interest, and if so, which (Chapter 2)
b. How can interests be measured? (Chapter 3)
c. How can scores for these scales be developed and what is the meaning of these scores? (Chapter 4)
d. What are the psychometric qualities of the scales and which differences in score can be interpreted? (Chapter 5 and 6)

We shall summarise the answers to each of these questions.
a. In a preliminary investigation with 176 items amongst 390 boys and girls, a cluster analysis revealed 17 different statistical clusters, all of which were comprehensible in terms of content and could be interpreted as distinct fields of interest. From this it may, we believe, be concluded that children of 11–13 years of age already have a systematic preference for certain fields of interest. This preference even appeared to be already characterized by a fairly subtle degree of differentiation. After a second preliminary investigation with 291 subjects, a test with 18 fields of interest, each consisting of 10 items, became available.

b. In chapter 3 the two forms of items usually employed in interest tests, namely those of the scale system and those of the choice system, were further analysed and their influence examined.

In addition to a number of incidental drawbacks, each of the item forms appeared to have one disturbing characteristic which is of great influence upon the meaning of the resulting interest scores and their susceptibility for interpretation. In *choice items* the subject must choose from a number of occupations and activities a fixed number which he prefers (eg. 1 activity out of 2, or 1 out of 4, or 6 out of 18, etc.) whereby all the occupations or activities chosen are also scored. This item form leads to equal sums of scores for all subjects. A set of scores in which this is the case is termed ipsative. Apart from the fact that as a result of the

1. This summary is, with a few alterations, a translation of chapters 7 and 14.
2. These scales are also termed empirical scales; since, of course, the construction of homogeneous scales is also based upon empirical data, we prefer the term „differentiating"

ipsative character of the scores the inter-correlations, as well as correlations with other data, are considerably reduced, the chief difficulty is that inter-individual comparison of ipsative scores is in principle impossible. It is, therefore, never possible to conclude on the basis of interest scores based upon choice items, that one person has more interest for a particular field than another.

Moreover, the intra-individual comparison of scores based upon choice items is in practice impossible. For in fact each item is measured by a different scale of measurement, namely, the set of items with which it is successively combined for the purpose of choice. Scores based on different scales of measurement are, of course, not mutually comparable.

With *scale items*, whereby the subject must give his degree of preference for each individual item according to a fixed scale, intra-individual comparison of scores is meaningful. For it may then be assumed that the interpretation of the given scale by one individual within one test situation is subject to little change. On the other hand, inter-individual comparison is not possible as a result of individual differences in interpretation and use of the given scale of preference (response set). It is concluded, especially on the basis of the possibility of individual comparison of the scores with scale items, that the scale form must be considered the most desirable for interest items.

c. In chapter 4 these insights are used to arrive at a system of scoring for the homogeneous scales of the ISI interest test.

The five scale positions (like very much, like, indifferent, dislike, dislike intensely) are scored from 5 – 1 inclusive, after which a raw score can be calculated for each of the 18 fields of interest by adding the item scores. These 18 scores together yield an interest-profile in which level, scatter and shape may be distinguished. The differences in level and scatter must be chiefly attributed to differences in two forms of response set: the intensity and the variability with which the subjects express their interest, which finds expression in the level and scatter of the profile respectively.

Since the response set is not for us a relevant datum, it can be eliminated from the scores. Differences in intensity of expression of interest can be eliminated by assimilating all profiles to one level. The result is then, however, that the sum of the new scores is the same for all subjects, so that these new scores are then also ipsative, with all the drawbacks associated with this. The form of the profile is not, however, impaired by ipsatising the raw scores. The differences of scatter could be eliminated by assimilating all profiles to one another with regard to scatter. This has not been done, because then incidental errors in profiles with a small original scatter may be greatly magnified, and in this way lead to erroneous interpretations.

Before reaching a definite decision with regard to the use of raw or ipsative scores, the consequences of ipsatising were further investigated. It appeared that the variance of the ipsative scores was always lower than that of the corresponding raw scores, and also that this loss of variance was not equally great for all scores. A comparison of factor analyses of the raw and ipsative scores showed that the first general factor of the raw scores is lost by ipsatising. It was remarkable that the loadings of the diverse fields of interest in the general factor varied, and did so in accordance with the differences in loss of variance by ipsatisation. This renders it probable that the general factor is also determined by an interest factor, which is manifest in several fields of interest, as well as by the response set. We are here concerned with the technical-manual interests, which by virtue of their relative over-representation in the 18 fields of interest, are partially incorporated in the general factor, only to be erroneously eliminated by the ipsatisation. In an interest test with a more balanced representation of the diverse fields of interests, the first general factor will be determined more exclusively by the response set.

On these grounds it was decided to choose the ipsatised scale-scores as the final score form for the fields of interest, despite the above-mentioned drawback. They have the same characteristics as the raw scores for intra-individual comparisons. Since, however, the intensity with which interest was expressed has been eliminated from the ipsative scores, inter-individual comparisons are also to a certain extent possible, namely in the

sense that it may be concluded that a certain field of interest occupies a higher place in the scale of preferences with one person than it does with another. Group comparisons are also made possible by this, which is the reason why ipsative scores can also be given in a standardised form.

Alongside these score forms directly deducible from the raw scores, factor scores and scores of the various profile characteristics (level and scatter of the raw score profile and scatter of the standard score profile) have also been developed.

d. In chapter 5 the item analysis of Form I of the interest test is discussed. The test satisfies the requirements with regard to degree of popularity and the item-total correlations, whilst the subdivision of the items in fields of interest on the basis of cluster analysis is satisfactory.

Finally, in chapter 6 the reliability and possibility of profile interpretation is discussed. The split-half reliability coefficients of the raw scores lie on the same level as in the school achievement test and intelligence test of the ISI series. Those of the ipsative scores, however, are distinctly lower. This is quite understandable from the more limited variance of the ipsative scores: the standard errors of measurement of raw and ipsative scores are, namely, similar to one another and of such a magnitude that a difference of 10 points between two scores is of minimal significance at the 6,8 % level.

From an analysis of the stability coefficients it appears that the ipsative scores are relatively more reliable than the raw scores: differences in profile-level between two testings influence all scores of fields of interest in the same way, so that the form of the profile is relatively more stable than might be expected on the basis of differences in the raw scores.

Chapter 6 closes with the question as to whether it is quite right to ignore individual differences in variability of expression of interest in interpretation of the profile: for subjects who frequently employ the extreme positions of the scale of preferences will exhibit significant differences sooner than subjects employing the more intermediate positions, even when they are consistent in their reactions to the items of the various fields of interest. When one starts out from the same standard errors of measurement in profile interpretation in both cases, this difference in variability of expression of interest is not taken into account, and it may be wondered whether wrong conclusions are not drawn in some cases. This question receives further examination in the second part of this study.

This second part deals with the construction and characteristics of the second version of the ISI interest test. The difference from the first version lies in the fact that alongside the homogeneous scales, differentiating scales have also been constructed for Form II, and moreover that there are only 8 instead of 18 homogeneous scales. Each factor of Form I has, namely, been reduced to a single scale, whilst in addition the social interests have been incorporated in the test as the eighth scale. Finally, the scoring system of the homogeneous scales has been changed.

The possibility of reducing the number of homogeneous scales and including the social interests was examined in a preliminary investigation, reported in chapter 8. The result of this preliminary investigation was a test with 155 items distributed over 8 fields of interest, each with 18–22 items. This test was given in the standardization research for Form II of the ISI-series, after which each of the homogeneous scales was reduced to 16 items on the basis of the item analysis of the data of this investigation.

In order now also to take the difference in variability of expression of interest into account in the profile interpretation, (not done in Form I, because there profile interpretation is based upon the theory of reliability), a totally new system of interpretation was designed. The starting point was taken in the item scores which, by means of giving tests on forms which are subsequently punched by an optical page-reader, became directly available for electronic treatment.

Interpretation of a profile is permitted when the average item scores of the 8 groups of items differ significantly from one another. This can be checked by means of a simple analysis of variance on the item scores

of each individual subject. If the analysis of variance is significant (which was the case at the 5 % level in a good 90 % of the standardization sample of Form II) differences found in average item scores of the 8 fields of interest do not appear to be based on chance and they can therefore be interpreted.

Finally, on the basis of the item variance within the 8 groups of items, a critical difference can be calculated by means of which each of the separate differences between two averages can be checked. Every difference which is greater than this critical difference is significant at the 5 % level and can therefore be interpreted.

Starting out from this critical difference, which varies from person to person, significance scores with the following characteristics have now been developed for the homogeneous scales: *a.* the differences of level between individuals are eliminated, these scores are therefore ipsative too; *b.* part of the differences in variance between persons, namely the differences in item variance within the fields of interest is eliminated; *c.* a difference of 10 points between 2 scores of one individual is significant at the 5 % level.

It was noticeable that the stability coefficients of the significance scores were somewhat higher than those of the raw scores. By eliminating the differences of level and part of the differences of variance, the information regarding the shape of the profile appeared to become relatively more stable. Since the form of the profile contained the most important information, and partially on account of the characteristics of the significance scores mentioned above, these are chosen as final score forms for Form II; for the purpose of group comparisons, possible with some reservations with the significance scores, as in the case of the ipsative scores from Form I, standardised significance scores are also given. The most important profile characteristics are also summarised in a score in Form II.

The remaining chapters of part II give an account of the construction of the differentiating scales. The principle underlying this kind of scale is stated in chapter 10. Taking test data as their starting point, scales may be developed which facilitate a maximal differentiation between 2 or more groups of persons. A score on such a scale indicates the extent to which an individual is in the same way as other members of a particular group distinguishable from another group. The problems which arose in the homogeneous scales with regard to differences of intensity of interests in one individual, or between several individuals do not arise here, because no pronouncement is made concerning the intensity of interest but about the degree of similarity with the interests of a particular group.

In the differentiating scales, we have drawn a distinction between classifying and contrasting scales. In *classifying scales*, various groups are differentiated with regard to the same reference group: in this way we have constructed scales which differentiate as well as possible between pupils from the V.W.O., M.A.V.O., L.N.O., L.A.V.O. or L.T.B.O.[1] on the one hand, and a reference group consisting of all first year pupils of secondary education on the other, according to interests. In *contrasting scales*, 2 groups are contrasted in order to make it possible to achieve optimal differentiation between them: in this way we have for instance developed a scale for distinguishing as well as possible between L.N.O. and L.A.V.O.[1] pupils according to their interests.

Description of the criterion groups which yielded the empirical data for the scale construction, and justification of the choice of reference group, is followed in chapter II by an exposition of why ipsatised item scores rather than raw item scores are taken as the starting point for scale construction. If we should start out from the raw scores, the difference in response set would be discounted in the scales in an unverifiable manner and so obscure their meaning.

A measure for the differentiating capacity of the separate items was found in the F-value of the difference in average ipsative item score of criterion- and reference-group. It was shown that when all F-values of the separate items are calculated on the basis of equally large groups, this F-value and the percentage over-

1. Different forms of the Dutch school system.

lapping traditionally taken as the measure for the differentiating capacity of a score, are functions of each other. All items with an F-value above a certain minimum were now taken up in the scales. On the basis of the size of the F-values and the direction of the difference between criterion- and reference-group, item weights of -5 to $+5$ were ultimately assigned to the items of each scale. The raw scale scores are calculated by adding the products of the item weights and the ipsative item scores.

The standardization of the raw scores is done differently in the classifying scales from in the contrasting scales. In the first case they are scaled by standardization of the raw scores of the *criterion groups* to a mean of 10,5 and a standard deviation of 2. When a standard score is 10 or more, that is to say higher than a half standard deviation below the mean of the criterion group, it can be said that the person is distinguishable from the group of reference with respect to his interests in the same way as the members of the criterion group.

The contrasting scales are scaled by standardization of the raw scores of both contrasting groups together towards a mean of 10 and a standard deviation of 3. A high score signifies greater resemblance to the one group with regard to interests, a low score to the other.

The capacity of the scales to differentiate is described in chapter 12, both for the criterion groups with regard to which they have been calculated and for a number of control groups separately. Most classifying scales appear to have good classifying possibilities, despite clear differences. These differences are in general clearly in accordance with the degree of agreement and difference between the types of formal education distinguished.

The contrasting scales also appear to differentiate well in general. The scales for distinguishing between girls and boys on the basis of interests work particularly well.

In order to investigate the quality of differentiating scales in yet another way, the extent to which it was possible to assign children from the first class of secondary education to the type of school actually attended was explored. Using fixed criteria, this appeared to be possible in about 45 % of the cases, whilst the 33 % assigned to the wrong type of school were by and large allocated to types of school closely resembling their own.

Cross validation showed the differentiating ability of the scales in new groups to differ very little from that in the original groups. Finally, the stability coefficients of the differentiating scales appeared to be on the same level as those of the ipsative scores of the homogeneous scales.

In the concluding chapter the intercorrelations between the various differentiating scales were analysed. Partially on the basis of these it was decided to incorporate all the classifying and only three of the contrasting scales in the final scoring system.

In view of the lack of data pertaining to predictive validity, the differentiating scales are put forward for experimental purposes. Whilst they appear to possess the differentiating capacity envisaged, it is not known in how far they can predict school success, etc. too. For this reason they are at present only suitable for purposes of further scientific research.